解答例

STEP II 1994、第 13 問

確率・統計 ・ クーポン収集問題

> **STEP II 1994、問 13**
>
> 先に進む前に、まず問題を読んでおこう。問題は STEP Support Programme の問題データベースにある：
>
> step.maths.org/questions/94-s2-q13

この問題は、*Beyond STEP* の第 9.11 節「**全種類を集める**」で詳しく解説されている：
https://beyondstep.keithmcnulty.org/ja/ch09/coupon-collector.html

鍵となる考え方

これは古典的な「クーポン収集問題」（coupon collector problem）であり、データベースという分類が付いているものの、中身は純粋な確率の問題である。全種類がそろうまでの長い待ち時間を、n 個の短い待ち時間に分割する。すなわち、1 枚目の新しいカードを得るまでの待ち時間、2 枚目を得るまでの待ち時間、という具合である。それぞれの短い待ち時間は幾何分布（geometric distribution）に従うので平均は容易に求まり、期待値の線形性（linearity of expectation）によってそれらを足し合わせることができる。合計は n と調和和（harmonic sum）の積になり、積分の不等式はその和の大きさを知るために用意されている。

解答

新しいカードを待つ

各袋には n 種類のカードのうち 1 枚が入っており、どのカードも確率 $\frac{1}{n}$ で等しく現れ、他の袋とは独立（independent）であるとモデル化する。

すでに r 種類の異なるカードを持っているとする。ただし $0 \le r \le n-1$ である。1 つの袋から新しいカードが出る確率は

$$p = \frac{n-r}{n} > 0,$$

であり、すでに持っているカードが出る確率は $q = 1 - p = \frac{r}{n}$ である。最初の新しいカードが出る袋までを含めて、開ける袋の数を X とする。袋どうしは独立であるから、$X = k$ とは、既出のカードが $k-1$ 回続いた後に新しいカードが出ることを意味する。よって

$$\mathrm{P}(X = k) = q^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

したがって

$$\mathrm{E}(X) = \sum_{k=1}^{\infty} kpq^{k-1} = p\,(1-q)^{-2} = \frac{p}{p^2} = \frac{1}{p},$$

となる。ここで、$|q| < 1$ のとき成り立つ $\sum_{k\geq 1} kq^{k-1} = (1-q)^{-2}$ を用いた（等比級数（geometric series）を微分するか、$(1-q)^{-2}$ を二項展開すればよい）。$r \leq n-1$ であるから $0 \leq q < 1$ である。ゆえに

$$\mathrm{E}(X) = \frac{n}{n-r},$$

となり、題意が示された。

（より手際のよい方法もある。最初の袋から新しいカードが出ればそこで終わり、そうでなければ確率 q で振り出しに戻る。よって $\mathrm{E}(X) = 1 + q\,\mathrm{E}(X)$ となり、$\mathrm{E}(X) = 1/(1-q) = 1/p$ を得る。ただしこの方法は $\mathrm{E}(X)$ が有限であることを前提としており、それは上の級数によって確かめられる。）

積分の不等式

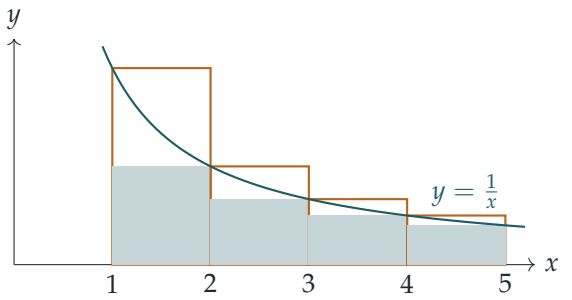
$r \geq 1$ とする。$x > 0$ において $1/x$ は減少関数であるから、$r \leq x \leq r+1$ に対して

$$\frac{1}{r+1} \leq \frac{1}{x} \leq \frac{1}{r},$$

が成り立つ。長さ 1 の区間で積分してもこの不等式は保たれるので

$$\frac{1}{r+1} \leq \int_r^{r+1} \frac{1}{x}\,\mathrm{d}x \leq \frac{1}{r}.$$

図で見ると、$x = r$ から $x = r+1$ までの $y = 1/x$ の下の面積は、高さ $\frac{1}{r+1}$ の低い長方形と高さ $\frac{1}{r}$ の高い長方形の間にある。



（見やすさのため、縦方向を拡大してある。）塗りつぶした長方形は曲線の下にあり、輪郭だけの長方形は曲線の上にはみ出している。

ここで $n \geq 2$ として、$r = 1, 2, \ldots, n-1$ についてこの不等式を足し合わせる。積分はつながって 1 つになる：

$$\sum_{r=1}^{n-1} \frac{1}{r+1} \leq \int_1^n \frac{1}{x}\,\mathrm{d}x \leq \sum_{r=1}^{n-1} \frac{1}{r}.$$

中央の項は $\ln n - \ln 1 = \ln n$ であり、左辺の和では $r+1$ を改めて r と書き直すと、r は 2 から n まで動く。よって

$$\sum_{r=2}^{n} \frac{1}{r} \leq \ln n \leq \sum_{r=1}^{n-1} \frac{1}{r},$$

となり、題意が示された。

全種類を集める

ちょうど r 種類の異なるカードを持っている間に開ける袋の数を X_r とする。ただし、$r+1$ 種類目のカードが出た袋までを含める。全種類をそろえるのに必要な袋の総数は

$$T = X_0 + X_1 + \cdots + X_{n-1}.$$

である。

> **考えてみよう** X_r はそれ以前に起きたことに依存する。それでも単純に期待値を足し合わせてよいのだろうか。

よい。期待値の線形性 $\mathrm{E}(A+B) = \mathrm{E}(A) + \mathrm{E}(B)$ には、独立性はまったく必要ない。（実際、それまでの経過がどうであれ、X_r は $p = \frac{n-r}{n}$ の幾何分布に従うので、前半の結果がそれぞれの X_r に適用できる。）したがって、$k = n - r$ とおくと

$$\mathrm{E}(T) = \sum_{r=0}^{n-1} \frac{n}{n-r} = n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{1}\right) = n\sum_{k=1}^{n} \frac{1}{k},$$

となる。

$H_n = \sum_{k=1}^{n} \frac{1}{k}$ とおく。不等式の左右それぞれから

$$H_n = 1 + \sum_{k=2}^{n} \frac{1}{k} \leq 1 + \ln n, \qquad H_n = \sum_{k=1}^{n-1} \frac{1}{k} + \frac{1}{n} \geq \ln n + \frac{1}{n}.$$

が得られる。よって

$$n\ln n + 1 \leq \mathrm{E}(T) \leq n\ln n + n.$$

上界と下界の差は n 未満であり、n が大きいとき $n\ln n$ に比べてごく小さい（両者の比 $1/\ln n \to 0$）。したがって、n が大きいとき、食べることになる袋の数はおよそ

$$n\ln n \text{ 袋}$$

と見込まれる。たとえば数学者が 100 人なら、Cruncho をおよそ 460 袋食べることになる。

検算

- $r = 0$ のとき：最初の袋は必ず新しいカードであり、$n/(n-0) = 1$ と一致する。$r = n-1$ のとき：特定の 1 種類だけが欠けており、それは確率 $\frac{1}{n}$ で現れるので、期待値は n 袋であり、$n/1$ と一致する。
- $n = 2$ のとき：$\mathrm{E}(T) = 2(1+\frac{1}{2}) = 3$ であり、下界と上界はそれぞれ $2\ln 2 + 1 \approx 2.39$、$2\ln 2 + 2 \approx 3.39$ である。
- $n = 20$ で全種類の収集を 10^5 回シミュレーションしたところ、平均は 71.9 袋となり、$20H_{20} \approx 71.95$ とよく合った。下界と上界は 60.9 と 79.9 である。

よくある間違い

- 積分の不等式を $r = 0$ から足し始めること：そこでは上界 $\frac{1}{r}$ が意味をなさず、$\int_0^1 \frac{1}{x}\,\mathrm{d}x$ は発散する。和は $r = 1$ から始めなければならない。
- 期待値を足し合わせるには X_r が独立でなければならないと考えること。線形性に独立性は不要である。
- 答えを $\ln n$ や n としてしまうこと：袋の数の期待値は $n\ln n$ のように増加し、n より速く増える。

STEP のその先へ

差 $H_n - \ln n$ は減少しながらある極限に収束（converge）する。これがオイラー（Euler）＝マスケローニ（Mascheroni）定数（Euler–Mascheroni constant）$\gamma \approx 0.5772$ である。したがって、より精密な答えは $\mathrm{E}(T) \approx n(\ln n + \gamma)$ であり、$n = 100$ なら 460 袋ではなくおよそ 519 袋となる。このように和と積分を比較する方法は級数の積分判定法（integral test）であり、大学の解析学（analysis）の最初の講義で早い段階に登場する。クーポン収集問題そのものも、アルゴリズムの解析や、トランプの山を何回ランダムにシャッフルすれば十分に混ざるかという研究の中に再び現れる。Grimmett と Welsh による *Probability: An Introduction*（確率論入門）は、幾何分布に従う待ち時間と期待値の線形性が組み合わせて使われる様子を見るのに適した本である。

Beyond STEP 解答集：https://beyondstep.keithmcnulty.org/ja/。解答本文は CC0 1.0 である。問題そのものは © Cambridge University Press & Assessment に帰属し、ここには掲載していない。英語版から機械翻訳したものであり、まだ校閲を経ていない。